GHAJAR
EXHIBIT 48

Page 1

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

----------------------------------x

RICHARD KADREY, et al.,   :

                Plaintiffs,   :

  v.   :   Case No.: 3:23-CV-03417-VC

META PLATFORMS, INC.,   :

                Defendant.   :

----------------------------------x

    HIGHLY CONFIDENTIAL, ATTORNEYS' EYES ONLY

            Videotaped Deposition of

                HUGO TOUVRON

        Tuesday, December 3, 2024

            10:06 a.m. GST

            London, England

Reported by: Cassidy Western, RPR, CRR

_____

            DIGITAL EVIDENCE GROUP

        1730 M. Street, NW, Suite 812

            Washington, D.C. 20036

                (202) 232-0646

12/3/2024   Richard Kadrey, et al. v. Meta Platforms, Inc.   Hugo Touvron
Highly Confidential - Attorneys' Eyes Only

Page 12

```
 1                    * * * * *

 2                    HUGO TOUVRON,

 3    of lawful age, being first duly sworn or affirmed

 4      to testify to the truth, the whole truth, and

 5    nothing but the truth, was examined and testified

 6                    as follows:

 7       EXAMINATION BY COUNSEL FOR THE PLAINTIFFS,

 8                 RICHARD KADREY, et al.

 9    BY MR. SIMONS:

10        Q    And with that, good morning,

11    Mr. Touvron.  My name is David Simons.  I

12    represent the plaintiffs.  I'm going to be the

13    person sort of asking you the questions today.

14             Could you just state your name for the

15    record.

16        A    (In English) Hugo Touvron.

17        Q    And are you presently employed?

18        A    Yeah.  I'm present employee at Meta.

19        Q    At Meta Platforms?

20        A    Yeah.

21        Q    And what is your current title?

22        A    Research -- AI research scientist.
```

12/3/2024   Richard Kadrey, et al. v. Meta Platforms, Inc.   Hugo Touvron
                    Highly Confidential - Attorneys' Eyes Only

Page 124

1   maybe it can be tested with some experiment. But

2   otherwise, I would say it's hard to say.

3   Q   And you're not sure if Meta's done that

4   type of experiment?

5   A   No. Yeah, I'm not sure we did that

6   because it can be quite complex to do this kind of

7   experiment. If -- yeah.

8   Q   Is there a reason why in LLaMA 1, for

9   instance, books were used -- there was a high

10   proportion of epochs for books than there was for

11   other datasets?

12   A   I think it's also relative to the size

13   of the dataset because, first of all, we try to

14   adapt -- if I remember correctly, we try to adapt

15   the sampling of our data in order to match what

16   was in the Chinchilla paper. The Chinchilla paper

17   from Google.

18   And we try to adjust our weighting

19   according to that. And I think this is the

20   rationale of why we adjust this sampling and why

21   we have these epoch. But it's a long time ago,

22   and I don't remember exactly what was the

Page 125

1   discussion and the rationale to do all this

2   adaptation.

3   Q   In your experience, are books useful for

4   long context training specifically?

5   A   I don't -- I don't run long context

6   experiment. Yeah, so I don't know -- I don't know

7   if it's useful. You mean for at pretraining time?

8   Q   Yeah, let's start with pretraining.

9   A   It's -- yeah, for me, it's not clear if

10   it's useful at pretraining time. Yeah, I mean, we

11   need to do some experiment. But there is some pro

12   and there is also some -- some cons related to how

13   the pretraining is designed.

14   Q   If the LLaMA models didn't train on

15   books or scientific articles, in your opinion,

16   would that degrade their performance?

17   A   It depend of the knowledge. So, for

18   instance, if there is a particular knowledge that

19   is not present in any other dataset, yes. If this

20   knowledge is present in other dataset, maybe not.

21   Because there is -- and it depend on the knowledge

22   because there is some knowledge you can infer from

Page 256

```
 1   guess, prompts and responses that have been used

 2   to -- to ensure that the model doesn't output

 3   copyrighted material?

 4        A    You mean if we are fine-tuning the model

 5   on some particular answer to avoid the model to

 6   provide -- well, I mean --

 7        Q    Yeah, that --

 8        A    If it -- if it has a particular question

 9   to provide the model to answer something?

10        Q    That would be one example, yeah.

11        A    So for instance, if you ask the model

12   how to make a bomb, he should not answer that.

13        Q    Well, that.  But apply it to copyright.

14   Like if I -- if you ask the model to give me the

15   entire copy of "Moby-Dick," has there been any

16   training put into the model to avoid the

17   possibility that the model will output an entire

18   copy of "Moby-Dick"?

19        A    For me, it's not super clear for -- for

20   instance, for "Moby-Dick" because I don't know if

21   there is a copyright issue with "Moby-Dick."  I

22   mean, I am --
```

Page 257

 1   Q   Sorry --

 2   A   -- completely not familiar with --

 3   Q   No, that's fine. Let's do one like

 4   "Harry Potter," a more recent novel. Is there any

 5   copyright mitigation that you understand, any

 6   training that's been done to avoid fine-tuning

 7   or -- you know, any fine-tuning or any -- any sort

 8   of that sort that's been done to the model so that

 9   if someone asks for it to output the entire copy

10   of "Harry Potter" that it won't do that? It will

11   refuse to do that?

12   A   From what I remember, we put particular

13   data in order to have some templated answer for

14   some particular question like "make a bomb," for

15   instance. But I don't know exactly if it's for

16   what is called copyright in a legal sense.

17   Because, for instance, when I say "copyright," I'm

18   pretty sure my definition is not the legal one.

19   Q   So let's just say if someone were to ask

20   the model to provide a copy of "Harry Potter" and

21   the answer the model provides is I'm sorry, I'm

22   not able to do that because the book is under

Page 258

1   copyright, how would it know to provide that

2   answer?

3   A   What we did is we incorporate some --

4   some data in order to have this kind of templated

5   answer, and this is part of the safety work,

6   provide some particular templated answer for some

7   particular question.

8   Q   And when you say we incorporate this

9   data to have this kind of completed answer, what

10   is the data that you incorporated? How does that

11   process work?

12   A   I think it's some human annotator that

13   annotates that. I'm not sure 100 percent. Maybe

14   some generated (indiscernible). I don't know.

15   But I think it's maybe human annotator.

16   Q   And do you know -- and then the model is

17   trained on these annotations to know that if

18   somebody asks for a particular copy of work, to

19   provide that type of response?

20   A   I'm sorry. Can you repeat?

21   Q   Yeah, so then you referenced annotated

22   data, the model would be trained on that annotated

Page 450

```
 1   CERTIFICATE OF SHORTHAND REPORTER - NOTARY PUBLIC

 2        I, CASSIDY WESTERN, RPR, CRR, the officer

 3   before whom the foregoing deposition was taken, do

 4   hereby certify that the foregoing transcript is a

 5   true and correct record of the testimony given;

 6   that said testimony was taken by me

 7   stenographically and thereafter reduced to

 8   typewriting under my direction; that reading and

 9   signing was requested; and that I am neither

10   counsel for, related to, nor employed by any of

11   the parties to this case and have no interest,

12   financial or otherwise, in its outcome.

13        IN WITNESS WHEREOF, I have hereunto set

14   my hand this 6th day of December, 2024.

15   My commission expires August 4, 2025.

16

17

18

19

20   _____        _____12/6/24____

21   Cassidy Western, RPR, CRR                    Date

22
```